School of Computing, Engineering & Digital Technologies

Teesside University

Middlesbrough TS1 3BA

Detecting Payment Fraud in Finance Industry using Machine Learning based Techniques

An academic research paper for possible submission to

IEEE Big Data Mining and Analytics Journal

Submitted in partial requirements for the degree of MSc Data Science

Date: 27 August 2021

Yu Shing Lui – A0201746

Supervisor: Dr. Qiang Guo (Larry)

Acknowledgements

During the writing of this master project, I have received many advices and assistance. Needless to say firstly, I would like to thank my project supervisor Dr. Qiang Guo (Larry) for his invaluable supervision and tutelage.

I would also like to thank my lecturer, Alessandro Di Stefano for his advice, encouragement. Most importantly, his inspiration on me for this project.

Last but not least, I would like to thank my parents and my partner for giving me their huge encouragement. I would not have completed this project without their support.

Thank you

Table of Contents

Detecting Payment Fraud in Finance Industry using Machine Learning based Techniques

Author

YU SHING LUI
School of Computing, Engineering & Digital Technologies
Teesside University
Middlesbrough, TS1 3AB
a0201746@tees.ac.uk

Supervisor

Dr. QIANG GUO (LARRY)
School of Computing, Engineering & Digital Technologies
Teesside University
Middlesbrough, TS1 3AB
Q.Guo@tees.ac.uk

1 Abstract

Card payment is one of the most popular payment methods to use for online purchasing, and it is particularly so during the recent pandemic period. Nevertheless at the same time, large amount of payment fraud transactions also exists thus causes many financial companies and card users in suffering financial losses every year. Identifying payment fraud transaction is always a big challenge as these transactions are complicated. Since the last couple of years, machine learning technique is getting popular to apply in different situations as it can solve various problems in a wide range of industries. As such, this project attempts to apply machine learning algorithms approach to detect payment fraud transactions which also endeavour to find out the best algorithm. The dataset of payment transactions come from European card holders, and it has around 600,000 transactions. Python has been used for the programming language to analyse the dataset. It mainly focuses on the algorithms of Random Forest, Logistic Regression, K-Nearest Neighbor, Naive Bayes, and Support Vector Machine. The result compares the accuracy rate which evaluate the performance between those algorithms. The result of them is 99%, 93%, 99%, 91%, and 94% respectively. The computational cost of each algorithm obtained are 1 minute 4 seconds, 8 seconds, 1 minutes 38 seconds, 0.4 seconds, and 11 hours 3 minutes 29 seconds respectively. From the result, Random Forest and K-Nearest Neighbor algorithms are the highest accuracy rate, but Random Forest computational cost is lower than K-Nearest Neighbor.

Keywords – payment fraud transaction, random forest, logistic regression, k-nearest neighbor, naive bayes, and support vector machine

2 Introduction

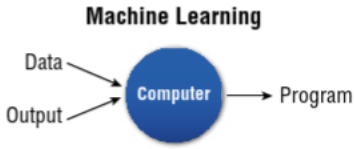
Since the onset of the pandemic incident, customers have been changing their spending pattern from offline to online. Especially during the lockdown period, most people have to buy groceries from online shop, even people could go shopping in supermarket. Also, almost of shops accepted credit or debit card for payment only in most time. The data from office for national statistics [1] in United Kingdom mentioned that the lockdown period from November 2020 to December 2020, it increased the index from 85.83 to 103.52. By then, customers are getting used to pay by card for shopping in daily life and it is the reason why the concern in payment fraud transactions has been more pressing as compared to the last couple of years.

Generally, payment transaction has two ways to make a payment. The first channel is offline payment [2]. Customers arrive at shops or restaurants (the merchant) and present their credit or debit cards for payment once they have purchased the merchandise or have rendered the service. Subsequently, the card issuing bank (For instance, NatWest Bank, Barclays Bank, or Lloyds Bank etcertica.) sent this transaction for approval. The network for card payment is linking between the card issuing company (For example, Visa, Master, or American Express etcertica.) and the card issuing bank. Once the company of credit or debit card issuer approves the transaction, such payment will be paid to the bank. On the other hand, the second channel is online payment [2]. Customers can do shopping from the websites, and they can input their payment card details to make a payment through the website's online payment system, like PayPal. If the payment approves, the merchant will be able to obtain the money back from an online payment system.

Payment fraud transaction has been increasing as now fraudsters can get a customer's card details more easily from a third-party hacked computers or they can get it through illegal channels such as "dark web". They can even steal it from merchant's websites or servers. Credit card fraud is one of the most well-known cybercrimes [3] and it is not easy come up on how much these financial companies actually have lost through cybercrime every year. In 2018, an American global computer security software company, McAfee, joined the Center for Strategic and International Studies (CSIS) which has an estimation of around $445 billion to $600 billion United States Dollar worldwide relating to cybercrimes [4].

Detecting payment fraud transactions have various approaches which using anomaly detection models and machine learning technologies are more popular ways in many financial companies nowadays [3]. The model of anomaly detection can find, avoid, and block payment fraud transactions occurred. It can reduce the number of payment fraud transactions occur and minimize the amount of financial lose from companies and credit or debit card users. In this project, it applies different models and experiment for payment fraud detection. In addition, the dataset for this project occupies large amount of normal payment transaction versus insignificant amount of fraudulent transactions. High accuracy rate is vital as credit card fraud creates huge financial loss in the banking sector. Therefore, it is aimed at detecting even the slightest suspicious transactions while machine learning is one of the most effective ways to solve this problem [5].

Machine learning is a common term which means data scientist uses an algorithm to apply into a dataset and predict the result or forecast the consequence based on a large amount of data [6]. Machine learning applies using the statistics of theory to create mathematical models. Data scientist has to use the model to train the data and find an efficient algorithm to resolve the situation in real life. The mission of machine learning requires a computer with outstanding processing power and therefore it has not been accomplished in the past few years owing to technological deficiency. At present, it is much better and even the cost of computer processing is lower than before. In figure 1, it is showed that computer is possible to finish the task of machine learning which using data and output to create the program [7].



The procedure of the program from data and output in machine learning
(Lee, 2019, pp. 3 [7])

Programming language is required for machine learning and Python is one of the most popular programming languages for data scientists to apply. Most of them will use scikit-learn as it is a popular library for Python programming language. It is including many useful models and algorithms with fantastic execution, and they can apply API commonly for a programming platform [8]. This project uses Python programming language for executing data analysis and all the models and algorithms base on scikit-learn library which it has most of common packages. It can apply on programming software called Jupyter without any difficulties.

The motivation for this project is using machine learning techniques to detect payment fraud transitions and aim to provide solutions to the financial companies in avoiding such risk efficiently. The outcome of this project is to und erstand the sequence of operations in detecting payment transactions with different algorithms. It can also provide suggestion of high accuracy rate and low computational cost from the result of evaluation.

Section three for literature view provides the general ideas of payment fraud transactions, process to detect and predict fraud with machine learning algorithms, and the difficulties of payment fraud detection. Section four for methodology mentions the theory of each algorithm will used in this project and show the flow chart of methodology to understand the sequences. Section five for evaluations explains the dataset feature, pre-processing the data before starting data analysis, and compare the result of confusion matrix, accuracy cost, and computational cost. Section six for conclusion is summarizing the result in this project and the future study in exploring other techniques to deter payment fraud transactions.

3 Literature review

3.1 Payment fraud transactions

Payment fraud is very common because it does not have high risks to earn a large amount of money in the short period. Once the card holder discovers the fraud transactions, those transactions may have already happened a few weeks ago. A bank needs to take time to chase back the transactions and it may lose money at the end if it cannot stop this transaction [10].

Some people's wallets are lost or stolen which their personal identity cards and bank cards can be sold to criminals. It is one of the most popular ways that personal information is lost in such way. The techniques of credit card fraud are by copying someone credit card and discover the password from the end user. The offender is possible to withdraw money from ATM or make a lot of purchase transactions from internet [1]. At the end, card holders may incur huge financial loss from fraud transactions, which in turn may reduce their intensity to use card in the future. For that reason, financial institutions start to take action about fraud detection.

The detection of payment fraud is started by providing a set of bank or credit card transactions and determine whether the transactions are fraudulent or non-fraudulent. Normally it is classified into two classes of transactions, which namely genuine class and fraudulent class [9].

3.2 Feature engineering

Basically, a feature is an assessable attribute which is like an amount of transaction. The meaning of feature engineering is picking up these meaningful attributes to apply algorithm. To predict fraud, the fraud detect team will look for some aspects of features. It classifies for seven categories [18]. Traditional features are traditional field to predict fraud, for example location, email address, transactions, and orders etcetera. It is all coming from the customer who find the data from the receipt. Behavioural features are based on customer behaviour's description. For instance, the length of time the customer spends in the webpage, the length of time the customer spends making orders, and speed of transaction etcetera. The purpose of these features is finding out a fraudster activity to compare with a normal activity from customers. Real time features are the updated real fraud incidences. Most likely, it is a categorical data which provides the percentage of fraudulent by category, for example, email domain, country area, and autonomous system numbers etcetera. The main object of the features is extending a fresh market which the company do not have an existing data. It can prevent the machine learning models to see any adverse effects. Individual customer features can provide information about the behaviour of customers. It should be internet protocol address, billing address and expenses etcetera. Session tracking features are the other features from customers. The feature can get it from JavaScript, which is like cookies, how customer copy and paste the card number for checking out, and the password storage etcetera. It can classify the genuine customer behaviour. Entity features can define as the customer's machine, home address, city, and emails etcetera. It can aware the fraud detection team to know about fraudster's delivery location. Network derived features are the features of network level. It can point to network shape that improve a customer data. For instance, the user account partaking for a family in the same location [19].

3.3 Machine learning algorithm apply to payment fraud transactions

Machine learning is a part of Artificial intelligence (AI) [15] and a data scientist can use a computer to analysis a large number of data value. Then it can determine a decision according to the data which similar way a human does. It uses in many fields now, for instances stock market prediction, image recognition, social media post recommendations and payment fraud detection.

For the traditional methods to detect fraudulent payments, the rule of false positive is always mentioned [16]. The result is a lot of amount false positives that the company blocks many of genuine customers. For instance, it determines high risk locations and large amount of orders are fraud case. It will also lose the genuine customers who want to spend expenses in specific location for high value of transactions. Also, the issue of fixed outcomes is also invalid on some situation. The fraudsters will change a fraud behavior always. If the prices are increased, the average value of the orders are also increased. If the rules are no change, that means the original rules may become invalid. The rule is also based on binary results, that means it does not allow to amend the outcome that a payment on the high risk. Furthermore, rules-based method needs to expand the library as fraud behavior develop as fast. However, it will create many works for the team and make the system slower because it needs more time to review manually. Hackers will change their methods always to steal customer data to pretend genuine customer to make fraud payment. It will be more difficult to use the original approach to detect new way of fraud cases. Even though all those methods are still important of detecting fraud, now data scientist can focus more than that [17].

3.4 Process to detect and predict fraud with machine learning

From the dataset, the fraudulent transactions are hard to discover as they all look like the legitimate transactions. The value of features is similar, and it is difficult to clarify if we do not know the feature of “fraud” in the dataset. In addition, categorical values are one of the concerns which it cannot support for machine learning algorithms. If the features of data are categorical value, they have to amend the value of data from categorical to numerical before applying data analysis [14].

Predicting fraud with machine learning which data scientist needs to understand the following process. The first thing is they must know the needs of the company, the criteria of the project’s success and minimize the fraudulent payment within specific percentage etcetera. Then, they need to find the historical data about fraudulent payment. The features of time, value of transactions, location and chargeback report etcetera are also worth to analysis. After received the raw data, cleaning data is essential and takes time. It normally takes over 50 percent of time for a project, and it requires specific technical skills [20]. Once it is done, data scientist needs to create machine learning model and it can provide a recommendation to report the new transaction is fraudulent or not. For building the models, it requires to use the cleaned data to find out the character and the best predictors of the fraudulent transactions. All of the features may include amount of transactions, customer behaviour and locations etcetera. When the model is trained, new data is needed to send to the model and provide a recommendation based on the trained system and it should decide to block the transaction if it is fraudulent. The team member still needs to do a manual review all the time to check a suspicious transaction and fine the model more accurate later. Furthermore, the models keep looking for feedback from the production environment about chargebacks and retain it all the time to make sure that it can

detect new fraudulent patterns. Finally, data scientist awares the machine's features of machine processing power, connection speed and GPU capabilities can increase the accuracy definitely [21].

3.5 Real time fraud detection

Fraud detection is processed by large number of transactions and utilized machine learning models, however the results take times. Maybe it needs to discover after few weeks and months to find out fraudulent transactions, because it is a difficult task absolutely. Fraudsters have a chance to make a fraudulent transaction as much as possible until the transactions are blocked. Real time fraud detection is an implementation that it can run the models once the online transaction is made. The system is possible to seek for fraud transactions without any delay. Then, the system will send notification to the bank to let them know it is a fraud transaction. It can reduce manpower cost for fraud detection team to take action for next stages [22].

3.6 Machine learning suited to fraud detection

Machine learning can provide a result very fast, and it will affect the experience of buyers who complete checkout. It requires the processing time does not take too long. Machine learning can process hundreds of thousands of requests and do comparison with the outcomes which discover the best answer. It may take within a second in real-time [23]. At the same time, it can evaluate the customer behaviour and their activity. It can decide to block a payment automatically if it spots an anomaly. At the same time, machine learning desire to increase the volume of data from transaction and it will not put more pressure on the rule's library. It can improve the accuracy rate with larger dataset. The reason is the system has more transactions to give the models picks the differences and similarities between customer behaviour more efficiently and predict the transactions in the coming future. Besides, it can save cost for manpower as it can proves hundreds of thousands of payments in the short time and human do not. The cost of manpower is enormous compare with the cost of machine learning. Even it can process the repetitive and monotonous tasks 24/7, and the team can focus on escalating decisions only. Additionally, it can perform more effective than human to discover pattern and trends. The models are possible to learn from historical data and adapt to changes a pattern to detect fraud transactions. It can recognise suspicious customers even they have not done chargeback yet [24].

3.7 Ethical Issue

Using fraud detection techniques to forecast fraud and genuine customers may trigger ethical issue. A model may find out some customers are genuine, but they are not actually. In contrast, a model detects some customers are fraudulent when they are genuine conscientiously. All these detection errors should be minimized. In fact, the bank's point of view about the cost prediction, which the cost of looking for a genuine customer that is a fraudulent is higher than the cost of looking for a fraudulent customer that is a genuine honest. The bank will lose the chance to get profit in the second situation. However, the bank will lose both the value of the loan and interest as well in the first situation. As the banks desire to maximise the profit, they will reduce the cost of misclassification and less concern about classifying customers is fraudulent or not. It should be concern about the genuine customers who have the same features of the fraudulent customers [25].

3.8 Difficulties of payment fraud detection

The first challenge of payment card detection is about dataset. The sensitive financial transaction dataset is confidential for the company, and it is difficult to search the dataset with features as the reasons of customer privacy. Due to the challenge of dataset, it has the other challenge about finding the existing work from research database. As the result, experimental results and real-world dataset for academic researchers are the biggest problems for payment or credit card detection.

Fraud transactions dataset is imbalance normally which is the feature of the dataset is disproportionate distribution. In general, non-fraud transactions should occupy a huge proportion of the dataset and it is a majority classes. On the other hand, the fraud transactions belong to the small proportion of the dataset, and it becomes as minority classes [11]. It is also a challenge to train the model if the researchers using machine learning models. Even the result of accuracy rate is very high, but it does not mean the model can find the fraud transaction punctuality. One of the popular techniques to resolve the imbalance dataset, which is oversample the minority class. Synthetic Minority Oversampling Technique (SMOTE) is possible to boost minority cases in a dataset used for machine learning [12]. It will have more explanation as following section.

Cleaning and finding the overlap dataset are also tough challenge as a data scientist has to handle an unclean data and usually takes plenty of time to make a dataset without any noise before processing. Minimize a noise in the data can increase accuracy and generate better result [10].

4 Methodology

4.1 Supervised Learning

From the outlook of machine learning, there are two kinds of learning skills which are supervised and unsupervised [26]. Data scientists apply specific algorithm with inputs and the expected outputs that the algorithm can seek a way to provide the expected outputs according to an input. Specially, an algorithm has ability to determine an output depends on the input. It does not need any help and even it does not know the pattern before. Data scientist can make a dataset which contains the expected result, machine learning will solve a problem most likely. The task of supervised machine learning can detect credit card transaction fraudulent activities. The input should be a credit card transaction and the output are whether the transaction belong to fraudulent or not. The financial companies store the card holder's transaction and record any transactions once it is fraudulent.

4.2 Random Forest

Random forest can use in classification and regression problems, and it uses classification for this experiment. It is effective in many different kinds of datasets, and it can fix the problem about the disadvantage of decision tree. The reason is decision tree has high chance to overfit the training set. Basically, random forest is gathering many of decision trees like an ensemble method and every single tree has a bit different from each others [27]. Each classifier finds out the own decision and they vote the most common or average for final decision.

The model of random forest is able to prevent of overfitting for the result via using stringent calculation and it can retain the prediction ability of the trees [28]. Random forest is selecting the points of data to build a tree and choosing a feature for each split test.

Random forest is using ensemble learning technique, which based on bagging algorithm. It can generate a lot of trees on the subset of data and merge the result of all trees. It can reduce the chance of overfitting in decision trees and improve the accuracy rate. Besides, it is stable algorithm. It is very difficulties to impact the other trees even a new data input to the dataset. Also, it is not much impacted by noise comparatively [28].

4.3 Logistic Regression

Logistic Regression is one of classification algorithms which it can provide a prediction of the input data belonging for a particular class. It will represent the class 0 or 1 for the output rather than a number [29]. For this experiment, the model has to determine the dataset value is fraud or not and it uses binary model to predict the results.

Sigmoid means a non-linear functions of logistic regression. The logistic function (or called sigmoid function and the input to logistic function are represent as below [30].

$$f(s) = \frac{1}{1 + e^{-s}}$$
$$s = w_0 z_0 + w_1 z_1 + w_2 z_2 + \ldots + w_n z_n$$

Defining about a specific class classifier, it requires to calculate the input "s" that is the optative coefficients "w" and the result of vector "z". The result of unput data "s" should be multiply each component and add them all. As mentioned before, the logistic function converts linear value to a probability should be 1 or 0. As the result, it belongs to 1 if the value is over 0.5 and the value is 0 otherwise [30].

Logistic regression is easy to execute and provides good efficiency in most of cases, especially the dataset of features is linearly separable. It does not require high standard computer to run this algorithm and suit to train the model for the first time. In addition, it is not easy overfitting in a low dimensional dataset when the dataset includes a large amount of data. Furthermore, it allows to update the model to reflect new data which using stochastic gradient descent [30].

4.4 K-Nearest Neighbor

K-Nearest Neighbors is not a complicated algorithm. It defers the learning while the test instance is granted, but it can perform a complicated approaches [31]. In an n-dimensional space, one point is represented by one training tuple. The combinations of n-attributes mean how many training tuple in it. KNN does not need classified that no specific training before the tuple is come. Normally normalization is necessary for some situations if part of attributes with large values and the others does not. Data transformation will use for pre-processing about data normalization approaches [32]. If confirmed a test tuple and the training tuples of k-nearest discover from the training space by a particular measure. It calculates the distance between training and test tuple and the k-neatest training tuples are called KNN [32].

$$distance(X_p, X_q) = \sqrt{\sum_{i=1}^{n} (x_{qi} - x_{qi})^2}, X_p, X_q \in R^n$$

The Euclidean distance is shown as below equation, and it can apply for numeric attributes only [33]. The solutions are changed the value of 1 or 0 to two attribute values for nominal attributes. Missing value of attributes is common, and many ways can solve this problem. Finding out the value of k can reduce the lowest rate of error in all the training tuples. KNN finding the class label by voting system which is the most common class of the test tuple.

K-Nearest Neighbors no need to learn anything and does not originate any discriminative function in the training period. It saves the training dataset and learn it from making a real time prediction. Due to KNN does not require training before making predictions, anyone can be added the data any time and it will not impact the accuracy rate. Also, it is easy to executive for binary problems, and it does not need any extra efforts to adjust multi class [33].

4.5 Support Vector Machine

Support Vector Machine is a classification algorithm which it can use for data classification of linear. It can be visualized as a surface which defines a border contains many points of data. All the points are feature values to plot in multidimensional space. The purpose of Support Vector Machine provides a flat border (named as hyper-plane) and separate both sides of data points evenly [34]. The concept is defined as linearly separable, which the target classes is divided by a linear equation with training tuples for the dataset.

The linear discriminant equation is indicated the linear hyperplane that is shown as below. X is represented by training tuple and W is represented by the weight vector. The goal of SVM discovers the best hyperplane that maximize the margin the data points of different classes [33].

$$W = (w_1, w_2, \ldots, w_n)$$

$$X = (x_1, x_2, \ldots, x_n)$$

$$g(X) = W * X^T + b = 0$$

Maximum Margin Hyperplane make the best distance of separation within two classes. MMH is important as the line create the best separation and sum up the greatest data in the future. The reason is the point may touch the line near the boundary if the line has just slightly changed in the positions. The support vectors means each class has the points are the closest the MMH and there are at least one support vector for each class. The main characteristic of SVM is saving a classification model in compact way by the support vectors, even though it has an incredibly large amount of features [35].

Support Vector Machine do not happen about overfitting issue and normally it performs well when it has a clear indication of separating between classes. It can perform well in terms of memory when total number of samples is less than the number of dimensions. SVM has the other important advantages that it can process high dimensional data effectively and finding the separating hyperplane. It can classify the data properly between different groups [35].

4.6 Naive Bayes

Naive Bayes algorithm is based the foundational mathematical theory, called Bayesian methods, for mentioning the probability of cases, and how probability should be modified according to an additional information. Bayes classification is making use of training data to measure the probability of each class rely on the value of feature. The classifier can use for unlabelled data to inspect probabilities to forecast the similar class for the new feature. Bayesian method uses all of proof to amend the predictions. Once the features have small changes, it may have large impact for the result [35].

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}=\frac{P(A\cap B)}{P(B)}$$

The Bayes theorem's formula is shown as above which mentioned the relationship with dependent events. The sign P(A|B) means the probability of case A given that case B happened, and it is called conditional probability. Due to the probability of case A is dependent on the event of case B occurred [36]. Naive Bayes algorithm using Bayes theorem for classification. It is very simple, repaid, and effective algorithm and easy to get the forecasted probability for a prognostication.

Naive Bayes performs classification is great and does not require a large amount of training data to reckon the test data. So, it will have a short period of training. Naive Bayes is not sensitive about noisy attributes, and it can prevent the risk of overfitting happened [36]. Furthermore, it does not occupy so much computational power as it runs efficiently. It is a good start to run the dataset at the first time even the dataset is large or not.

4.7 Confusion Matrix

A confusion matrix usually is a table or figure, which is a classifier to state out the performance [47]. It is normally collected from the test dataset and compare each class with all other class and find out the number of samples are not classified. From the structure of the table, it can discover some key metrics which is significant part of machine learning [46].

From the table 1, each class represents one of the following:

True Positive (TP): It means a model detects some customers are fraudulent which needs to block. The ratio of sample classifies properly to all positive sample: TP / TP + FN

True Negative (TN): It means a model detects some customers are genuine which does not need to block.

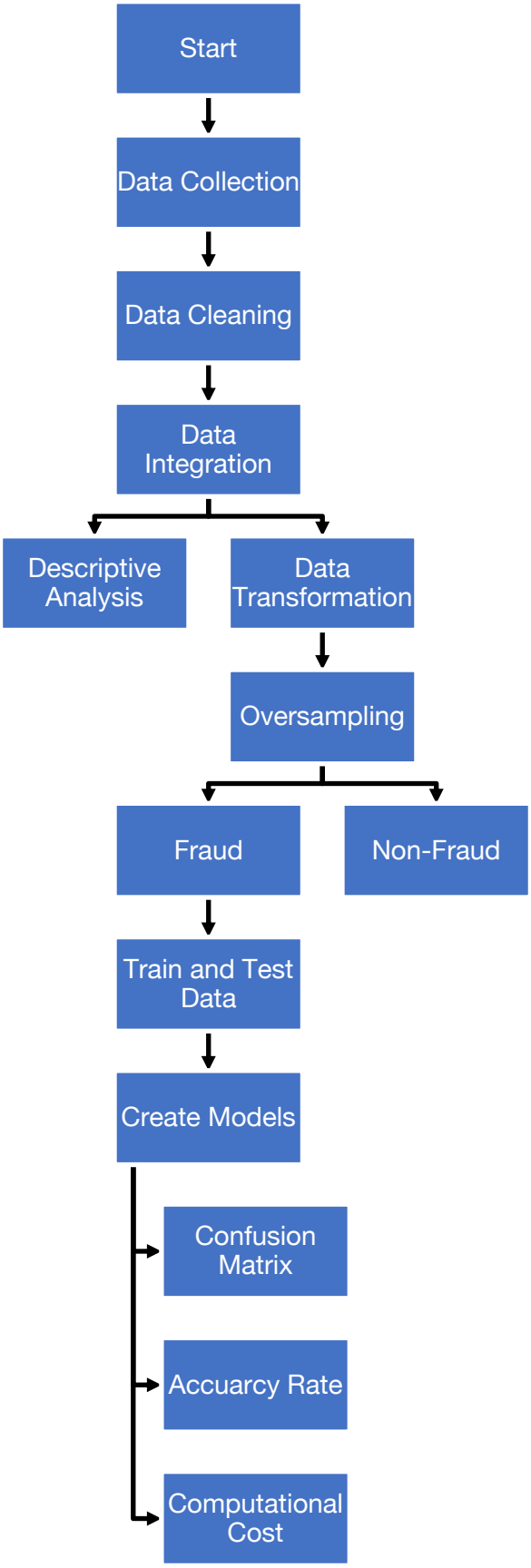
False Positive (FP): It means a model detects some customers are fraudulent, but they are genuine actually. It is also called Type I error. The ratio of sample classifies properly to all negative sample: FP / TN + FP

False Negative (FN): It means a model may find out come customers are genuine, but they are not actually. It is also called Type II error.

	c	
h	0	1
0	TN	FN
1	FP	TP

Table 1, how confusion matrix forms (Cichosz, 2015, p.196 [48])

4.8 The flowchart of methodology



5 Evaluations

5.1 Dataset description

For this project, it uses the same dataset from Lopez-Rojas [37] PhD studies. The dataset is a bank payment simulation. It is based the idea of on Multi Agent-Based Simulation and can be also called MABS. All of transactions about this dataset are come from a bank in Spain and the purpose is enhancing the big data of applications development. All transactions of data are started from November 2012 to April 2013. The location of zip code is come from Madrid and Barcelona. The dataset can report about the situations of payments, but it has not any privacy issue from the bank customers [38].

Due to the dataset's transactions are synthetic data which aggregated from a bank. It does not have any privacy issue, which like real transactions, personal information, and legal disclosure etcetera. So, it can use for education or other research. At the same time, the real data and the process of conversions have not been disclosured.

In this dataset [39], it has around six hundred thousand rows and ten columns. It includes the features of step, customer, age, gender, zipcodeOri, merchant, zipMerchant, category, amount, and fraud. The details of each feature will show as below, and it will understand how they process for classification.

1. Step

It represents the number of days for the transaction in this dataset. The range of this dataset contains around six months in total.

2. Customer

It can identify each bank customer with unique ID for their transaction. Although it is a unique identifier, it does not have any sensitive information for the customers. All customers unique code will convert to integer for analyzing in machine learning models.

3. Age

Each code indicates specific age range of each customer. It has eight categories to represent the customer age range and needs to change the value of string to integer format for further usage. In the table 2, it explains all of age range (below age 18 until over age 65) with the encoded code. For the last row of this table, it will clarify as unknown ages in this dataset.

Code	Age Range
0	<=18
1	19-25
2	26-35
3	36-45
4	46-55
5	56-65
6	>65
U	Unknown

Table 2, the table of age range with encoded code

4. Gender

Each code shows the gender type of each customer. There are four categories, which are enterprise, female, male and unknown column.

Code	Description
E	Enterprise
F	Female
M	Male
U	Unknown

Table 3, the table of the gender type with encoded code

5. Zip Code Origin

This column shows the zip code of each customer. However, it can find that all the zip codes are the same and it is no reason to use this feature. It will not have any impact for the result because of no difference of all of rows.

6. Merchant

It can identify each receiver with unique ID for their transaction. Although it is a unique identifier, it does not have any sensitive information for the merchant. All the merchant unique code will convert to integer for analyzing in machine learning models.

7. Zip Merchant

This column shows the zip merchant of each merchant. However, it is the same result of the feature of Zip Code Origin which all the zip codes are the same and it is no reason to use this feature. It will not have any impact for the result because of no difference of all of rows.

8. Category

This feature represents all categories for each transaction belongs to. In the table 4, it has the same situation as other features. The column of category needs to change the value of string to integer format for further usage. Otherwise, it cannot support for analyzing in machine learning models.

Category	Description
es_contents	Contents
es_food	Food
es_transportation	Transportation
es_fashion	Fashion
es_barsandrestaurants	Bars and restaurants
es_hyper	Hypermarkets
es_wellnessandbeauty	Wellness and beauty
es_tech	Technology
es_health	Health
es_home	Home
es_otherservices	Other services
es_hotelservices	Hotel services
es_sportsandtoys	Sports and toys
es_travel	Travel
es_leisure	Leisure

Table 4, the table of the description with each category

9. Amount

Due to the transactions of this dataset are from Spain, Europe. We believe that this feature of the amount should be in Euro and the values are numeric format with decimal values.

10. Fraud

It uses boolean format for this column of each transaction. The value of 1 means the transaction is fraud. Otherwise, the value of 0 means the transaction is non-fraud.

5.2 Pre-processing data

From the figure 1, it illustrates the transaction of fraud or non-fraud in this experiment. It has 594,643 transactions totally around six months period. According to the instance count, fraudulent and non-fraudulent transactions are unbalanced. It has only 7,200 transactions belonged as “Fraud” and it occupies around 1.2% in the dataset. Non-fraudulent transactions being almost imperceptible to the vision. At the same moment, it remains 587,443 transactions (98.8% approximately) are belonged as “Genuine”. This is a typical situation in fraud detection. It will affect the performance of classify model when imbalanced data is not processed beforehand. Most likely the predictions will rely to the major class and ignore the minor class in the attribute. The result will have high bias in the model finally [41]. Therefore, the imbalance dataset needs to add or remove data to make it balanced. Oversampling approach is considered in this situation. It is the process of creating synthetic data that attempts to create a sample of the feature from minority class randomly. One of the most common technique, Synthetic Minority Oversampling Technique (SMOTE) will use to oversample a dataset for solving this problem [40].

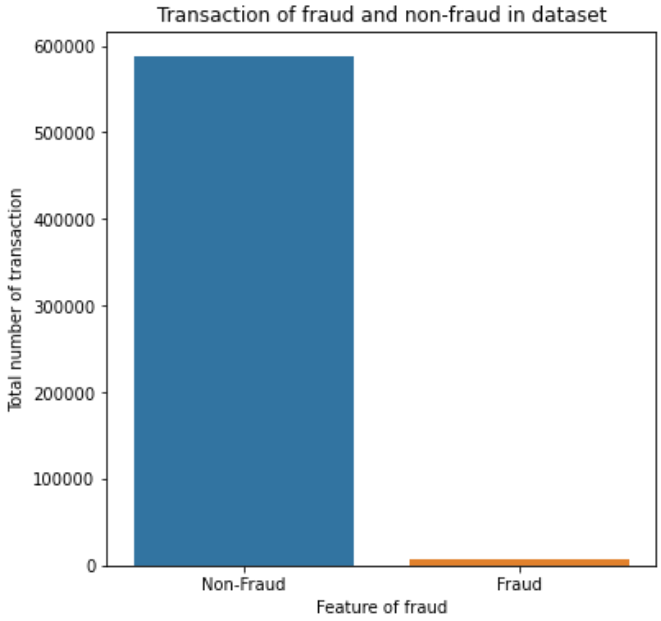


Figure 1, transaction of fraud and non-fraud in dataset

5.2.1 Min-Max Normalisation

Min-Max Normalisation can scale the difference by the range of the variables. For every attribute of the dataset, “0” should be the minimum value of the attribute and “1” should be the maximum value. The result of transformation will get into a decimal value between zero to one, unless there are some new data values which rely outside of the original range [45]. It notices that the feature of “step” and “amount” are wide range of values which compare with other features. As the result, min-max normalization can transform the value between 0 to 1 and reduce any inappropriate influence result when apply machine learning algorithms.

$$X^* = \frac{X - \min(X)}{\text{range}(X)} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

(Larose, 2004, p.36 [45])

At the below of table 5, it shows the result of finishing min-max normalisation. Before this process, the dataset will cross out the features of 'customer', 'zipcodeOri', and 'zipMerchant'. Those three eliminated variables are not very relevant for analyzing. In this dataset, it has some categorical features and need to transfer to numerical values. Otherwise, it cannot process afterwards. Besides, it noticed that the values of features are very wide range, and it will have inappropriate influence result when apply machine learning algorithms. Therefore, the values of features should normalise the variables. The purpose of normalization is to make all data in the same scale in every feature and Min-max normalisation is one of the most popular methods to do normalization [44].

	step	age	gender	merchant	category	amount
0	0.0	0.571429	0.666667	0.612245	0.857143	0.000546
1	0.0	0.285714	0.666667	0.612245	0.857143	0.004764
2	0.0	0.571429	0.333333	0.367347	0.857143	0.003228
3	0.0	0.428571	0.666667	0.612245	0.857143	0.002071
4	0.0	0.714286	0.666667	0.612245	0.857143	0.004288
...	...	...	...	...	...	...
594638	1.0	0.428571	0.333333	0.367347	0.857143	0.002465
594639	1.0	0.571429	0.333333	0.367347	0.857143	0.006090
594640	1.0	0.285714	0.333333	0.632653	0.142857	0.002694
594641	1.0	0.714286	0.666667	0.367347	0.857143	0.001736
594642	1.0	0.571429	0.333333	0.367347	0.857143	0.003233

Table 5, the result of finishing min-max normalization

5.2.2 Synthetic Minority Oversampling Technique

Synthetic Minority Oversampling Technique (SMOTE) is a statistical technique for adding the amount of data in the dataset until it is balanced. The module is possible to creating new cases from the original minority class in the dataset. In addition, the performance of SMOTE will not affect the amount of majority class at the same time [40]. Except the new cases copy of the existing minority class, it collects sample from the other class and neighbors which is closed to. Then, it generates new instances which merge attributes of the target case with attributes of the neighbors. The purpose of this approach is enhancing the attributes ready to every class and keep the sample more common [42]. This technique uses the whole dataset as an input, however it adjusts the percentage of the minority class only. For instance, the dataset in this project which the feature of fraud has 1.2 percent occupancy and it belongs to the minority class. This technique increases the percentage until it is balanced as majority class [43].

Once SMOTE is applied to balance the dataset, the results show that we have the exact number of class instances (1 and 0).

```
fraud
0         587443
1         587443
dtype: int64
```

Table 6, the result of SMOTE applied

5.3 Data Analysis

5.3.1 Confusion matrix

5.3.1.1 Random Forest Classifier

From the result of a confusion matrix (Figure 2) as below, the amount of true positive and true negative value is 175,880 and 174,213. Their amounts are very close. Also, it is a very high ratio compare with false positive and negative were only 353 and 2,020 that false negative is a bit more than false positive.

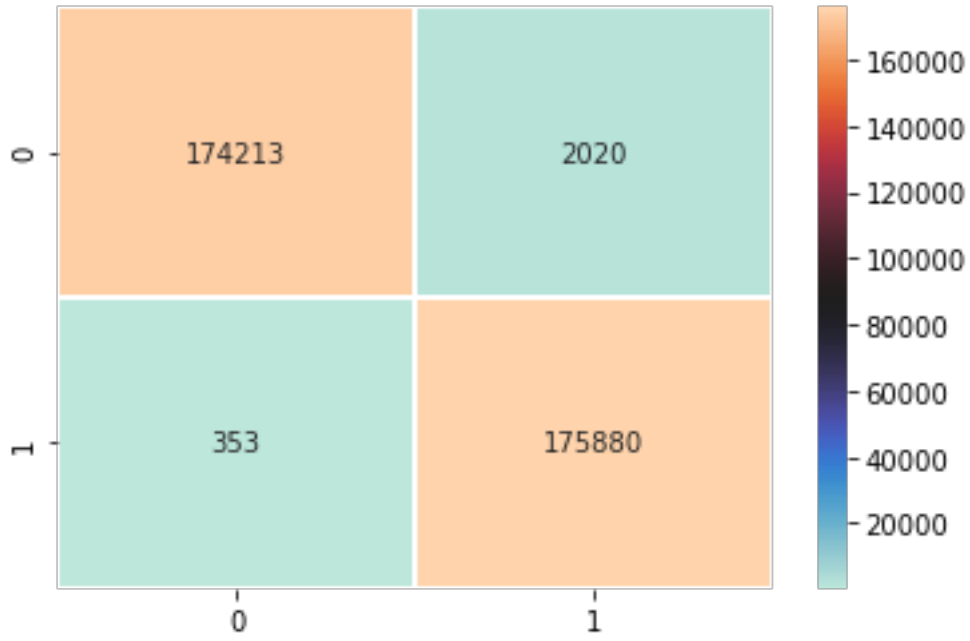


Figure 2. Confusion matrix heatmap of using random forest

5.3.1.2 Logistic Regression

From the figure 3, true positive and true negative values are 157,899 and 169,691 which true positive is bit less then true negative. It is a very high ratio compare with false positive and negative are 18,334 and 6,542. The amount of false positive is almost three times more than false negative in logistic regression.

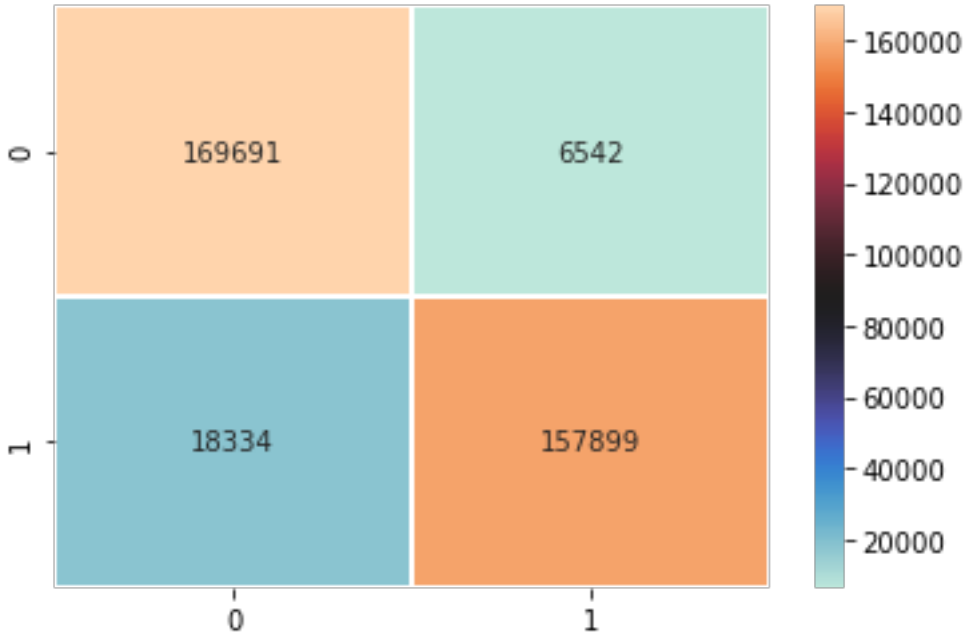


Figure 3. Confusion matrix heatmap of using logistic regression

5.3.1.3 K-nearest Neighbors Classifier

According to the result of a confusion matric (Figure 4) as below, true positive and true negative value are 175,848 and 173,533. This result is very similar as random forest. It is a very high ratio compare with false positive and negative are only 385 and 2,700.

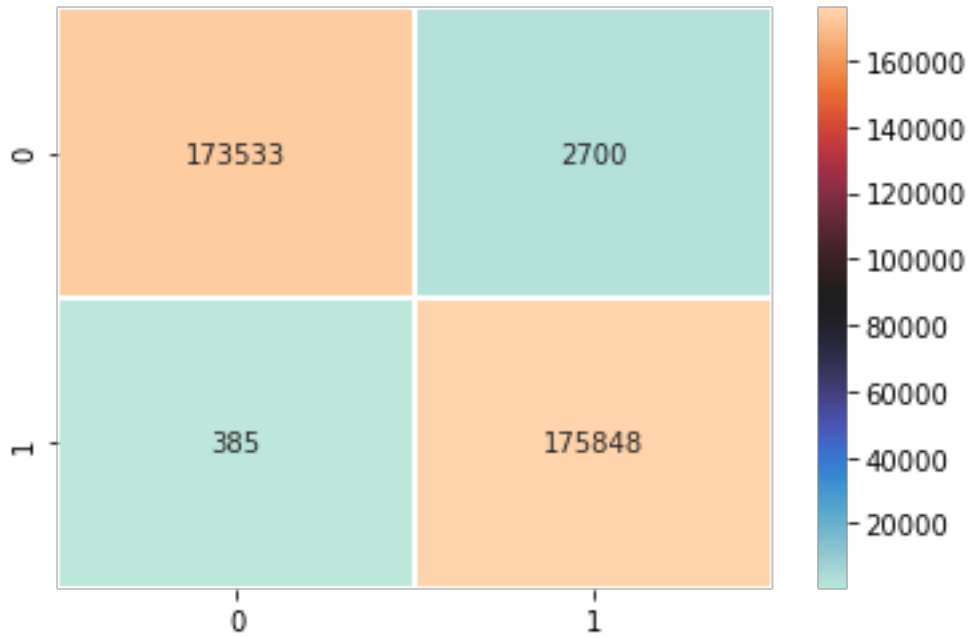


Figure 4. Confusion matrix heatmap of using k-nearest neighbors classifier

5.3.1.4 Gaussian Naïve Bayes

From the result of a confusion matric (Figure 5) as below, true positive and true negative value are 161,215 and 160,072. Both amounts are closed each others. Also, it is a very high ratio compare with false positive and negative were only 15,018 and 16,161. Their amounts are also very close dramatically.

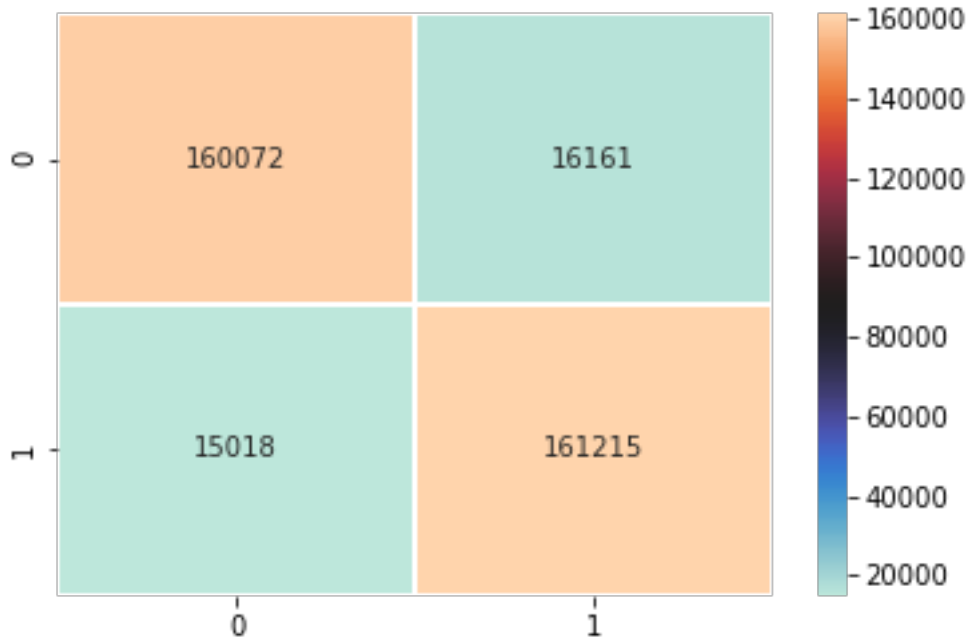


Figure 5. Confusion matrix heatmap of using gaussian naïve bayes

5.3.1.5 Support Vector Machine

From the result of a figure 6, true positive and true negative value are 159,492 and 170,368. The amount of true positive is a bit more than true negative. At the same time, it is a very high ratio compare with false positive and negative are only 16,741 and 5,865.

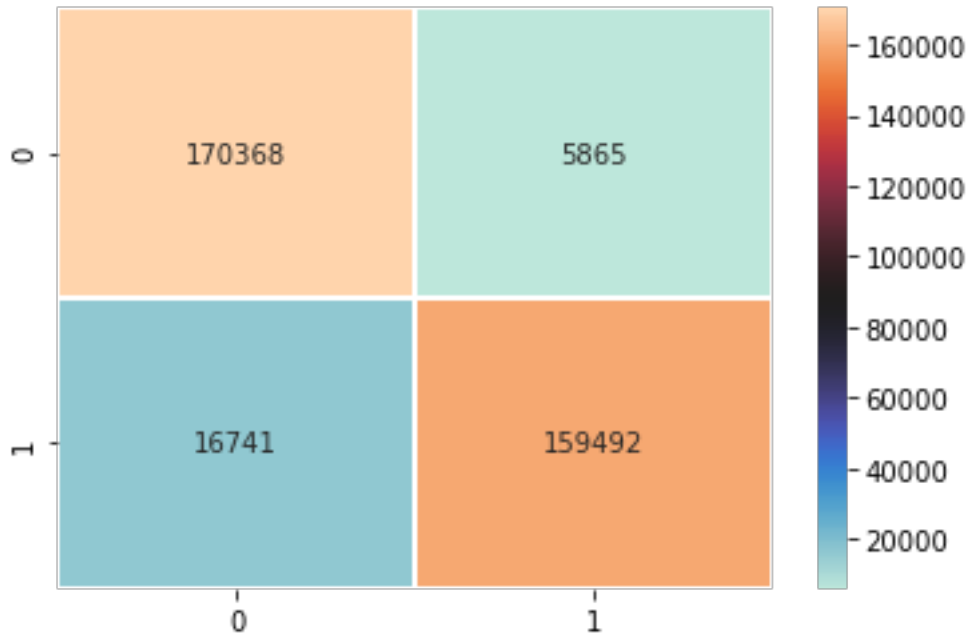


Figure 6. Confusion matrix heatmap of using support vector machine

5.3.2 Accuracy rate and computational cost

In the previous section, it mentioned the prediction in payment transactions by five different algorithms, which are Random Forest (RF), Logistic Regression (LR), K-Nearest Neighbor (KNN), Gaussian Naive Bayes (NB), and Support Vector Machine (SVM). In table 7, it showed the result of accuracy rate and the computational cost of each algorithm. In addition, it attempts to slices the size of dataset for six parts and compare the differences of each other. The result showed that the size of dataset is not affect the accuracy rate and it is very stable for each algorithm. The final accuracy rate obtained are 99%, 93%, 99%, 91%, and 94% respectively. However, the computational cost is varied, and it depends on the size of dataset. The computational cost of each algorithm obtained are 1 minute 35.4 seconds, 5.2 seconds, 2 minutes 42.6 seconds, 0.97 seconds, and 10 hours 30.1 minutes respectively. It is not difficult to find out SVM's processing time is incredibly high compare with the others, even it can perform very well accuracy rate. SVM may not be suitable for large dataset [49]. The reason is SVM must compute the data size at least quadratic, so the training complexity is very high [50].

	1-100,000 rows	**1-200,000 rows**	**1-300,000 rows**	**1-400,000 rows**	**1-500,000 rows**	**All rows**
Model	**Accuracy Rate**					
RF	99%	99%	99%	99%	99%	99%
LR	93%	93%	93%	93%	93%	93%
KNN	99%	99%	99%	99%	99%	99%
NB	92%	92%	92%	91%	91%	91%
SVM	94%	94%	94%	94%	94%	94%
Model	**Computational Cost**					
RF	8s	20s	36s	56s	1m11s	1m4s
LR	1s	3s	3s	4s	5s	8s
KNN	3s	8s	55s	43s	1m11s	1m38s
NB	0.1s	0.1s	0.2s	03s	0.4s	0.4s
SVM	21m37s	1h8m37s	2h30m23s	4h25m36s	7h6m6s	11h3m29s

Table 7, the accuracy rate and computational cost in different models

In this project, it contributes the other concern about analysing fraud payment detection. Most of dataset of fraud transaction is imbalanced datasets, because most of transactions are non-fraud cases and it just have a small amount of fraud cases only. It may need to pre-process Synthetic Minority Oversampling Technique (SMOTE) before applying the models. From the other studies [13] [38], it has implemented the models that it includes k-nearest neighbor, and gaussian naive bayes. The results are 99% and 98% separately. To make comparison between their efficiency and it can make a conclude that random forest and k-nearest neighbor are very effective models for fraud payment detection.

6 Conclusion and future study

Due to the need in detecting payment fraud transaction in finance industry, this project discussed about different machine learning techniques and also comparing their results. Besides, finding the dataset for this project is also one of the most challenging issues we confronted. The reason is the dataset may involve features of customer personal information, that is the reason why this project is using the synthetic data from Kaggle. In this project, it uses some classification methods and the accuracy rate of them is over 90% separately. In addition, the confusion matrix can provide the details about the value of true positive and true negative. From the result of confusion matrix, the prediction from the Random Forest algorithm is more accurate than the others even the computational cost is not the lowest. At the same time, this project also finds out Support Vector Machine algorithm may not the best option for this dataset as the computational cost is relatively high.

For future studies, it can explore and analyse other techniques such as neural networks, unsupervised, reinforcement and deep learning. It is possible to identify other techniques to find the optimal result and apply on payment fraudulent detection.

7 References

[1] www.ons.gov.uk. (n.d.). UK spending on credit and debit cards - Office for National Statistics. [online] Available at: https://www.ons.gov.uk/economy/economicoutputandproductivity/output/datasets/ukspendingoncreditanddebitcards.

[2] Emerald, G (ed.) 2005, E-Commerce, Emerald Publishing Limited, Bradford.

[3] Hwang, YH 2018, C# Machine Learning Projects : Nine Real-World Projects to Build Robust and High-performing Machine Learning Models with C#, Packt Publishing, Limited, Birmingham.

[4] Laudon, K, & Traver, C 2020, E-Commerce 2020-2021: Business, Technology and Society, EBook, Global Edition, Pearson Education, Limited, Harlow.

[5] Bhattacharyya, DK, & Kalita, JK 2013, Network Anomaly Detection : A Machine Learning Perspective, CRC Press LLC, Philadelphia, PA.

[6] Alpaydin, E 2014, Introduction to Machine Learning, MIT Press, Cambridge.

[7] Lee, W 2019, Python Machine Learning, John Wiley & Sons, Incorporated, Newark.

[8] Hackeling, G 2014, Mastering Machine Learning with scikit-learn, Packt Publishing, Limited, Olton Birmingham.

[9] Sakharova, I. (2012). Payment card fraud: Challenges and solutions. 2012 IEEE International Conference on Intelligence and Security Informatics.

[10] ResearchGate. (n.d.). (PDF) Machine Learning Techniques for Fraud Detection.

[11] Shivanna, A., Ray, S., Alshouiliy, K. and Agrawal, D.P. (2020). Detection of Fraudulence in Credit Card Transactions using Machine Learning on Azure ML. [online] IEEE Xplore.

[12] Zhang, Z. and Huang, S. (2020). Credit Card Fraud Detection via Deep Learning Method Using Data Balance Tools. [online] IEEE Xplore.

[13] Shiguihara-Juárez, P. and Murrugarra-Llerena, N. (2018). A Bayesian Classifier Based on Constraints of Ordering of Variables for Fraud Detection. [online] IEEE Xplore.

[14] Thennakoon, A., Bhagyani, C., Premadasa, S., Mihiranga, S. and Kuruwitaarachchi, N. (2019). Real-time Credit Card Fraud Detection Using Machine Learning. [online] IEEE Xplore.

[15] Khatri, S., Arora, A. and Agrawal, A.P. (2020). Supervised Machine Learning Algorithms for Credit Card Fraud Detection: A Comparison. [online] IEEE Xplore.

[16] Jain, V., Agrawal, M. and Kumar, A. (2020). Performance Analysis of Machine Learning Algorithms in Credit Cards Fraud Detection. [online] IEEE Xplore.

[17] Tanouz, D., Subramanian, R.R., Eswar, D., Reddy, G.V.P., Kumar, A.R. and Praneeth, C.V.N.M. (2021). Credit Card Fraud Detection Using Machine Learning. [online] IEEE Xplore.

[18] Adepoju, O., Wosowei, J., lawte, S. and Jaiman, H. (2019). Comparative Evaluation of Credit Card Fraud Detection Using Machine Learning Techniques. [online] IEEE Xplore.

[19] Rai, A.K. and Dwivedi, R.K. (2020). Fraud Detection in Credit Card Data using Unsupervised Machine Learning Based Scheme. 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC).

[20] Mittal, S. and Tyagi, S. (2019). Performance Evaluation of Machine Learning Algorithms for Credit Card Fraud Detection. [online] IEEE Xplore.

[21] Benson Edwin Raj, S. and Annie Portia, A. (2011). Analysis on credit card fraud detection methods. [online] IEEE Xplore.

[22] Thennakoon, A., Bhagyani, C., Premadasa, S., Mihiranga, S. and Kuruwitaarachchi, N. (2019). Real-time Credit Card Fraud Detection Using Machine Learning. [online] IEEE Xplore.

[23] Banerjee, R., Bourla, G., Chen, S., Kashyap, M. and Purohit, S. (2018). Comparative Analysis of Machine Learning Algorithms through Credit Card Fraud Detection. [online] IEEE Xplore.

[24] Varmedja, D., Karanovic, M., Sladojevic, S., Arsenovic, M. and Anderla, A. (2019). Credit Card Fraud Detection - Machine Learning methods. 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH).

[25] Delamaire, P. (2009). Title Credit card fraud and detection techniques : a review. Banks and Bank Systems, [online] 4(2).

[26] Bhattacharyya, S., Jha, S., Tharakunnel, K. and Westland, J.C. (2011). Data mining for credit card fraud: A comparative study. Decision Support Systems, 50(3), pp.602–613.

[27] Breiman, L. (2001). Random Forests. Machine Learning, [online] 45(1), pp.5–32.

[28] Xuan, S., Liu, G., Li, Z., Zheng, L., Wang, S. and Jiang, C. (2018). Random forest for credit card fraud detection. [online] IEEE Xplore.

[29] Bhattacharyya, S., Jha, S., Tharakunnel, K. and Westland, J.C. (2011). Data mining for credit card fraud: A comparative study. Decision Support Systems, 50(3), pp.602–613.

[30] Nadim, A.H., Sayem, I.M., Mutsuddy, A. and Chowdhury, M.S. (2019). Analysis of Machine Learning Techniques for Credit Card Fraud Detection. 2019 International Conference on Machine Learning and Data Engineering (iCMLDE).

[31] Tran, T.C. and Dang, T.K. (2021). Machine Learning for Prediction of Imbalanced Data: Credit Fraud Detection. 2021 15th International Conference on Ubiquitous Information Management and Communication (IMCOM).

[32] Beckonert, O., E. Bollard, M., Ebbels, T.M.D., Keun, H.C., Antti, H., Holmes, E., Lindon, J.C. and Nicholson, J.K. (2003). NMR-based metabonomic toxicity classification: hierarchical cluster analysis and k-nearest-neighbour approaches. Analytica Chimica Acta, 490(1-2), pp.3–15.

[33] Makhabel, B 2015, Learning Data Mining with R, Packt Publishing, Limited, Olton Birmingham.

[34] Saheed, Y.K., Hambali, M.A., Arowolo, M.O. and Olasupo, Y.A. (2020). Application of GA Feature Selection on Naive Bayes, Random Forest and SVM for Credit Card Fraud Detection. 2020 International Conference on Decision Aid Sciences and Application (DASA).

[35] Lantz, B 2013, Machine Learning with R, Packt Publishing, Limited, Olton.

[36] Dinov, I.D. (2018). Probabilistic Learning: Classification Using Naive Bayes. Data Science and Predictive Analytics, pp.289–305.

[37] ResearchGate. (n.d.). (PDF) BankSim: A Bank Payment Simulation for Fraud Detection Research.

[38] Vidanelage, H.M.M.H., Tasnavijitvong, T., Suwimonsatein, P. and Meesad, P. (2019). Study on Machine Learning Techniques with Conventional Tools for Payment Fraud Detection. [online] IEEE Xplore.

[39] kaggle.com. (n.d.). Synthetic data from a financial payment system. [online] Available at: https://www.kaggle.com/ealaxi/banksim1 [Accessed 14 Aug. 2021].

[40] Baesens, B, Van, VV, & Verbeke, W 2015, Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques : A Guide to Data Science for Fraud Detection, John Wiley & Sons, Incorporated, Hoboken.

[41] Zhang, Z. and Huang, S. (2020). Credit Card Fraud Detection via Deep Learning Method Using Data Balance Tools. [online] IEEE Xplore.

[42] Giussani, A 2020, Applied Machine Learning with Python, EGEA Spa - Bocconi University Press, Chicago.

[43] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, pp.321–357.

[44] Larose, DT, & Larose, CD 2015, Data Mining and Predictive Analytics, John Wiley & Sons, Incorporated, New York.

[45] Larose, DT 2004, Discovering Knowledge in Data : An Introduction to Data Mining, John Wiley & Sons, Incorporated, Hoboken.

[46] Joshi, P, & Joshi, P 2017, Artificial Intelligence with Python, Packt Publishing, Limited, Birmingham.

[47] Tanouz, D., Subramanian, R.R., Eswar, D., Reddy, G.V.P., Kumar, A.R. and Praneeth, C.V.N.M. (2021). Credit Card Fraud Detection Using Machine Learning. [online] IEEE Xplore.

[48] Cichosz, P 2015, Data Mining Algorithms : Explained Using R, John Wiley & Sons, Incorporated, Somerset.

[49] Zhang, D., Bhandari, B. and Black, D. (2020). Credit Card Fraud Detection Using Weighted Support Vector Machine. Applied Mathematics, 11(12), pp.1275–1291.

[50] Zhang, J.-P., Li, Z.-W. and Yang, J. (2005). A parallel SVM training algorithm on large-scale classification problems. [online] IEEE Xplore.